NC Department of Health and Human Services
Program Integrity

Statistical Sampling Operational Manual

N.C. Department of Health and Human Services – Program Integrity
Statistical Sampling Operational Manual

Purpose: Provide the steps in determining the size of a sample needed for auditing purposes, and the process to follow in obtaining and extrapolating that sample.

Summary of Steps in Process:

I. Data Need – Identify the data that is needed and the time-frame it is for
II. Data Query – Produce the data from pre-established DRIVE queries, or request that the information be produced from a customized query
III. Data Transfer – Transfer the data to the statistician who is in charge of producing the sample
IV. Data Analysis Phase I – Analyze the produced data to determine the number of records, the total dollar amount, and the distribution of the data
V. Data Analysis Phase II – Based on the distribution of the data, decide how to segment the data. Once it is segmented, get the mean and standard deviation for each segment from SAS, and enter this information into RAT-STATS to determine the sample size needed for the specified confidence interval
VI. Sample Size Determination – Consult with the person requesting the sample to determine which sample size is the best given their time and staff constraints. Samples with a higher confidence interval and precision mean auditing more records. There is a tradeoff in accuracy and time that must be agreed upon
VII. Sample Selection – Once the size is agreed upon, use RAT-STATS to produce random numbers which are then used in SAS to pull a random sample from the original data from step II
VIII. Sample Submit – Provide the random sample to the requestor and maintain a copy for historical record. Document the process so that the sample can be matched if necessary
IX. Extrapolation Process – After the sample has been audited, the requestor will then enter the dollar amount that should have been paid for each record, and return the sample to the statistician who will then use RAT-STATS to determine how much should have been paid for all the records from the original dataset. This information will then be returned to the requestor. This process will also be documented for potential use in legal matters

Detail of Steps in Process

I. Data Need: Section chief/staff express a desire to obtain a sample for auditing purposes. The sample and purposes will be unique to each requesting section, with the overall process for obtaining the sample consistent as a whole. The data will be queried for a selected time period.

II. Data Query: If the section chief/staff usually obtains data from a query that has already been setup up in the DRIVE system, then he/she will run the data for the time period and provider of interest for the sample. If there is a need for a data query to be designed for this purpose, then the staff person in charge of generating DRIVE queries will be contacted to design the query and produce the data.

III. Data Transfer: Once the data has been successfully obtained, it will then be forwarded to the staff statistician in charge of producing the sample (Bradford Woodard). The data will then be analyzed to determine the appropriate sample size.

IV. Data Analysis – Phase I: The following example will detail the steps for determining sample size using SAS and RAT-STATS:

The following is the first SAS program run in order to determine what the data looks like. It calculates the total number of records in the database, along with the total amount paid for the selected time frame. The amount paid will be the variable used to determine the sample size. The output from this program will demonstrate how the data is distributed by the amount paid.

Based upon the distribution of the amount paid, a decision is made about how to break the data up into groups. The goal is to divide the data into smaller groups with smaller spread so that the needed sample will then be smaller. The following program breaks the data into two groupings to be used to determine the sample. It has 16,303 records with the amount paid totaling $665,050.44.

<u>It should be noted that the following example uses two strata. If there is only one stratum used, then simply follow the same steps and use only one stratum.</u>

SAS Program for Step 1:

```
data t1 t2;
   set library.data ;
   if  paid < 40 then output t1;
   if  paid >= 40 then output t2;
   delete;


ods select moments quanitles;
```

```
proc univariate data=t0;
   var paid;
   title1 'claims details all';
run;

ods select moments;
proc univariate data=t1;
   var paid;
   title1 'claims details < 40';
run;

ods select moments;
proc univariate data=t2;
   var paid;
   title1 'claims details >= 40';
run;
```

Program Output:

The output below will provide the detailed distribution from the data.

claims details All

The UNIVARIATE Procedure
Variable: PAID (PAID)

Moments

N	16303	Sum Weights	16303
Mean	40.7931326	Sum Observations	665050.44
Std Deviation	3.64236471	Variance	13.2668207
Skewness	-0.498859	Kurtosis	4.82754539
Uncorrected SS	27345766.5	Corrected SS	216275.711
Coeff Variation	8.9288674	Std Error Mean	0.02852658

Quantiles (Definition 5)

Quantile	Estimate
100% Max	52.08
99%	49.56
95%	46.02
90%	46.02

75% Q3	42.48
50% Median	40.92
25% Q1	38.94
10%	35.40
5%	35.40
1%	31.86
0% Min	3.54

claims details < 40
The UNIVARIATE Procedure
Variable: PAID (PAID)

Moments

N	7593	Sum Weights	7593
Mean	37.7373449	Sum Observations	286539.66
Std Deviation	2.34790663	Variance	5.51266555
Skewness	-4.2593084	Kurtosis	39.4592012
Uncorrected SS	10855098.1	Corrected SS	41852.1569
Coeff Variation	6.22170594	Std Error Mean	0.02694475

claims details >= 40
The UNIVARIATE Procedure
Variable: PAID (PAID)

Moments

N	8710	Sum Weights	8710
Mean	43.4570356	Sum Observations	378510.78
Std Deviation	2.18849604	Variance	4.78951492
Skewness	1.40066684	Kurtosis	1.83759346
Uncorrected SS	16490668.3	Corrected SS	41711.8855
Coeff Variation	5.03599938	Std Error Mean	0.02344967

V. Data Analysis – Phase II: Based on the distribution of the data, segment it in order to group the most data together to lower the amount of spread within segments. This is a process that will take some judgment and will vary for each dataset. Once this determination is made, the information is input into RAT-STATS to determine the sample size.

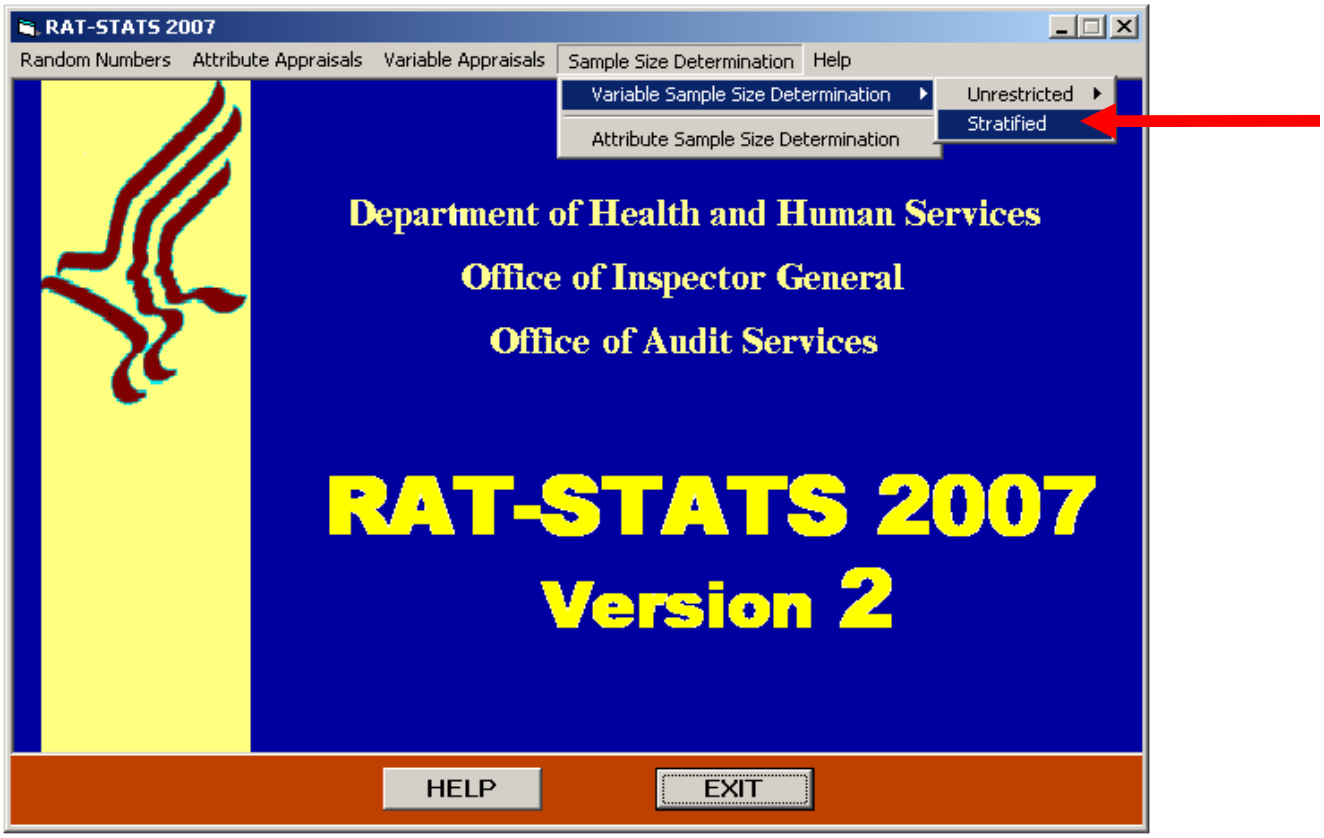
RAT-STATS Sample Size and Random Number Process:

The mean and standard deviation from each of the seven segments of data in the output above will be recorded for entry into RAT-STATS. An EXCEL spreadsheet will be setup to reference this information. An example EXCEL table is below:

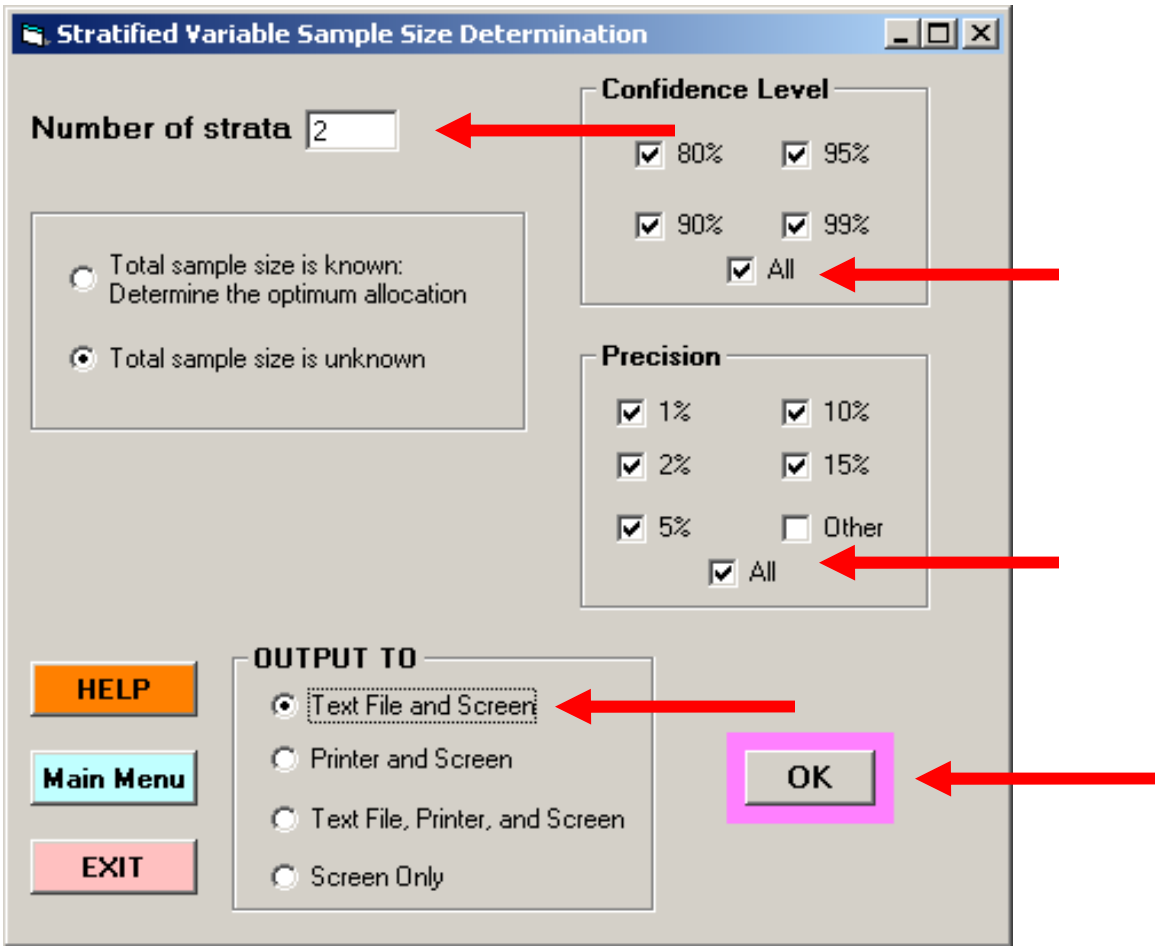
Strata	Mean	Std Dev	N	Sum	Sample	SMean	SSt.Dev.	SSum	Start	End	Start Num
s1 - <40	37.74	2.35	7,593	286,539.66	110	37.80	1.87	4,157.58	1	7,593	1
s2 - >=40	43.46	2.19	8,710	378,510.78	110	43.18	2.10	4,749.66	7,594	16,303	7,594
Sum	40.79	3.64	16,303	665,050.44	220	40.49	3.35	8,907.24			

The Sample numbers already appear here. In real time, they will be entered into the table after going through the RAT-STATS process documented next. These figures are not known at the start. The Start and End numbers are determined by the Sum of records in each Strata. These are used in RAT-STATS later on to determine where the random numbers start for each strata.

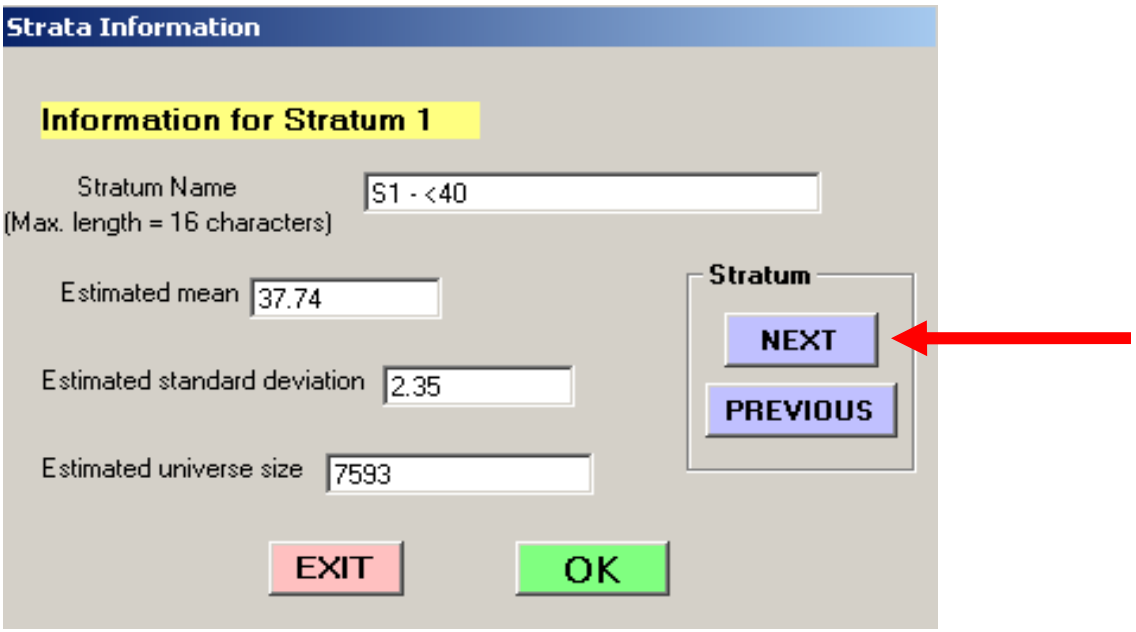
Once in RAT-STATS, select the Sample Size Determination, Variable Size Determination, Stratified button as shown here:



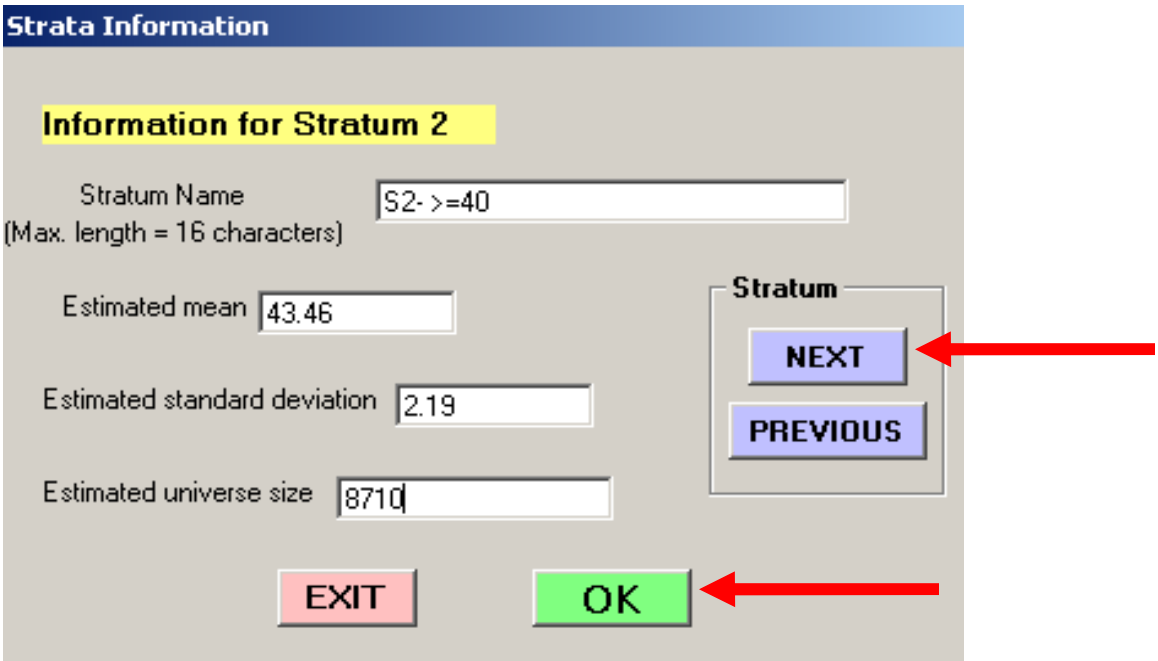
Next, enter the number of strata, 2 in our example. Check the All boxes for Confidence Level and Precision. Check the box with Total Sample Size is Unknown. Click the Output To Text File and Screen button. It will ask you where to save the file when you click it. Name it and save it to the correct location, then click OK.



Next, enter the name, mean, standard deviation, and universe size for the first strata (from the EXCEL table above) and then click Next.



Do the same thing for each of the strata and then click OK.



It will save the file to the location you indicated and display the results on the screen. The output will look like this:

Windows RAT-STATS
Statistical Software
Sample Size Determination

THE ESTIMATES ARE BASED ON THE FOLLOWING ENTRIES:

NBR	DESCRIPTION	-- MEAN --	-- STD.DEV. --	-- UNIVERSE --	-- RATIO --
1	S1 - <40	37.74	2.35	7,593	48.33%
2	S2 - >=40	43.46	2.19	8,710	51.67%
- TOTALS -		40.80	3.64	16,303	

= =

Sample Sizes for Stratum 1: S1 - <40

		Confidence Level			
		80%	90%	95%	99%
	1%	25 (*)	41	57	98
	2%	7 (*)	11 (*)	15 (*)	25 (*)
Precision	5%	1 (*)	2 (*)	3 (*)	4 (*)
Level	10%	1 (*)	1 (*)	1 (*)	1 (*)
	15%	1 (*)	1 (*)	1 (*)	1 (*)

Sample Sizes for Stratum 2: S2 - >=40

		Confidence Level			
		80%	90%	95%	99%
	1%	27 (*)	43	61	105
	2%	7 (*)	11 (*)	16 (*)	27 (*)
Precision	5%	2 (*)	2 (*)	3 (*)	5 (*)
Level	10%	1 (*)	1 (*)	1 (*)	2 (*)
	15%	1 (*)	1 (*)	1 (*)	1 (*)

Total Sample Sizes

		Confidence Level			
		80%	90%	95%	99%
	1%	52	84	118	203
	2%	14 (*)	22 (*)	31	52
Precision	5%	3 (*)	4 (*)	6 (*)	9 (*)
Level	10%	2 (*)	2 (*)	2 (*)	3 (*)
	15%	2 (*)	2 (*)	2 (*)	2 (*)

NOTE (*): One or more sample sizes were under 30. The generated sample sizes were the result of mathematical formulas and did not incorporate management decisions concerning the purpose of the sample or current organizational sampling policies. You may need to increase the sample sizes in order to be in compliance with organizational objectives.

It lists the appropriate sample sizes for each stratum and for the overall dataset.

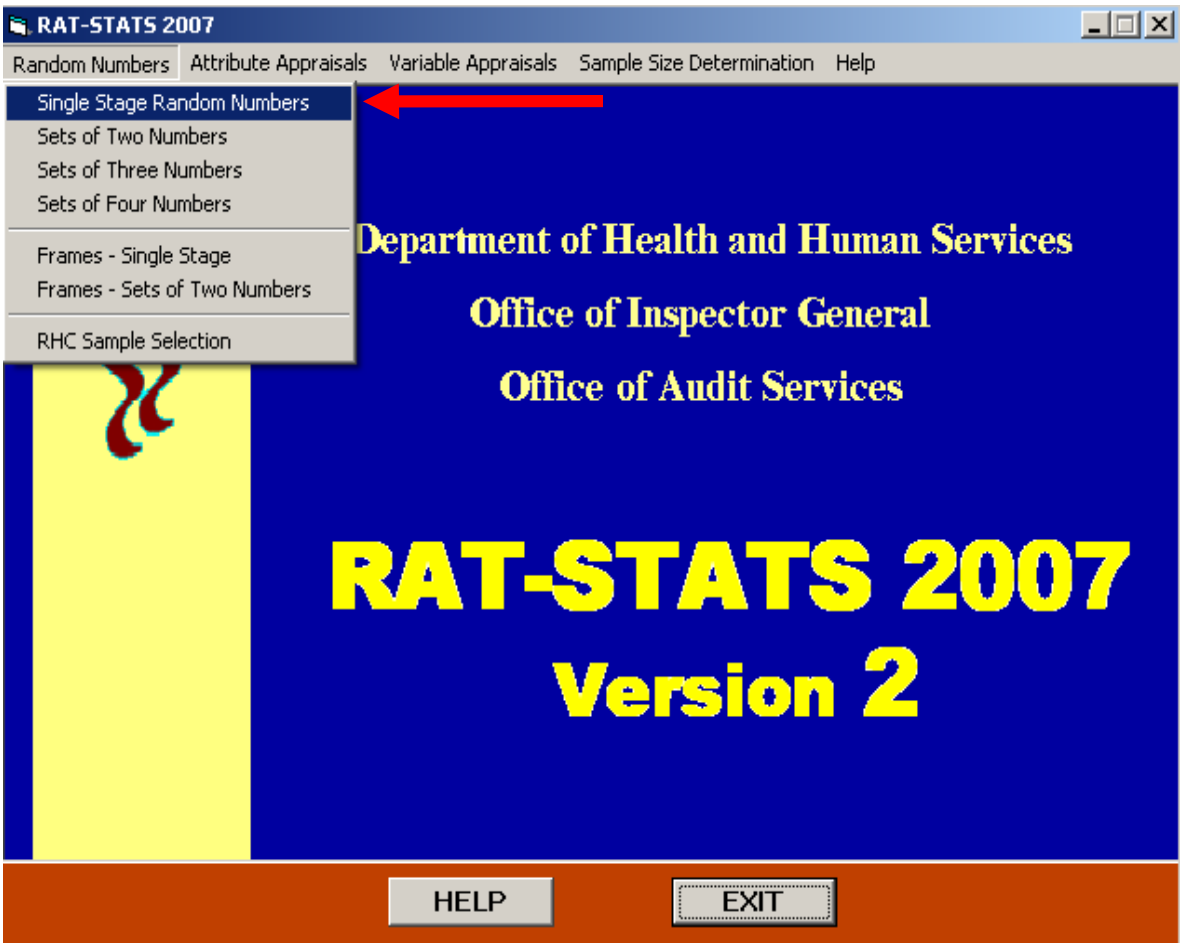
VI. Sample Size Determination: Program Integrity will utilize a minimum 95% confidence level and 5% precision level. Because there was little standard deviation within each stratum above, the number of recommended samples to obtain the minimum confidence and precision levels is small (3 for stratum 1, 3 for stratum 2, and a total of 6). Program Integrity must sample at least 100 total claim details for all samples, with at least 30 records from each stratum. This yields robust, sound statistical estimates. The RAT-STATS sample chart provides a guideline and is not absolutely necessary to use. It simply gives some helpful breakouts if needed.

In the above example, in order to obtain a sample with the maximum 99% confidence and 1% precision level, a sample of 220 records was selected. Selecting more samples than the recommended 203 is fine. If the desire is to have the maximum confidence and precision then a sample of at least 203 is needed. This gives a 99% confidence interval with 1% precision. This simply means that the sample will give a paid amount dollar figure that we can be 99% confident it will be within 1% of the actual dollar amount for the entire dataset.

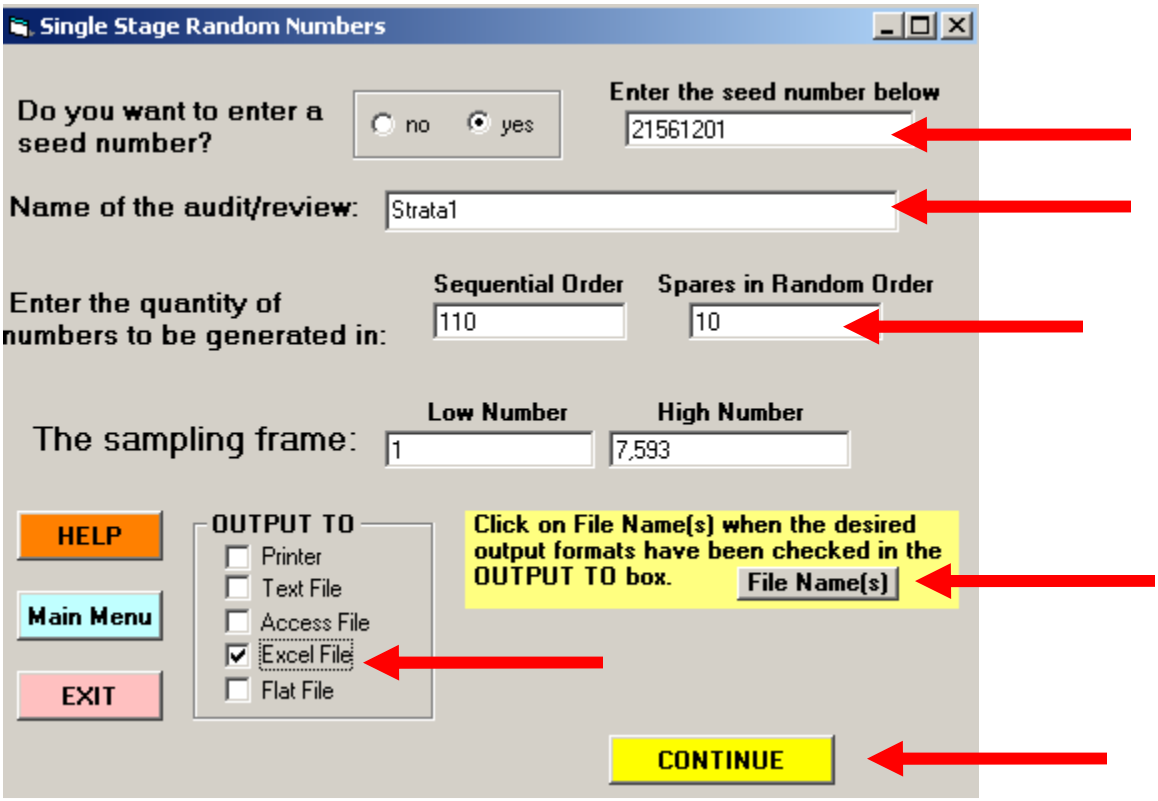
VII. Sample Selection: For example purposes, let's assume that the 99% confidence with 1% precision was sufficient for the requestor. This means that at least 203 records will need to be pulled from the 16,303 from the original dataset. The output above tells how many records are needed from each stratum. This information is then entered into the EXCEL spreadsheet from earlier, and used for reference in RAT-STATS to create random numbers. The table is provided again here:

Stratum	Mean	Std Dev	N	Sum	Sample	SMean	SSt.Dev.	SSum	Start	End	Start Num
s1 - <40	37.74	2.35	7,593	286,539.66	110	37.80	1.87	4,157.58	1	7,593	1
s2 - >=40	43.46	2.19	8,710	378,510.78	110	43.18	2.10	4,749.66	7,594	16,303	7,594
Sum	40.79	3.64	16,303	665,050.44	220	40.49	3.35	8,907.24			

The sample will need 110 random records from stratum 1, and 110 from stratum 2. Go back into RAT-STATS and select Random Numbers, Single Stage Random Numbers.



Next check the Yes button for the Do you want a seed number question. The seed number is a reference number that the program uses to produce random numbers. It allows for the same random numbers to be generated by someone else if they are trying to replicate the process. The number at the bottom of a dollar bill can be used as a reference seed number:



Name the strata, enter the quantity of numbers to be generated and spares in case they are needed. The sampling frame low/high numbers are where the strata begins/ends, as indicated by the reference table. Choose "Output To" EXCEL file, click the File Name button and save it to

the appropriate place. Then click Continue. The random numbers for this stratum will then be generated. This process must be repeated for each stratum, 2 in our example.

Once this is complete, there will be two EXCEL files with random numbers for each stratum. These two will need to be combined into 1 spreadsheet with all 220 numbers sorted from low to high. This will then be imported into SAS to be used to pull the random numbers from the dataset.

The following program is used in SAS to create the random sample:

```
data strata1;
    set library.data;
    if paid < 40;



data strata1a;
    set  strata1;
    length number 4;
    retain number 0;
    number = number + 1;
run;




 data strata2;
    set library.data;
    if paid >= 40 ;



data strata2a;
    set strata2 ;
    length number 4;
    retain number 7593;
    number = number + 1;


data all;
set strata1a strata2a;



proc sort data=all;
```

```
    by number;

proc sort data=library.randomnumbers;
    by number;

data sample.samptot;
    merge all(in=s) library.randomnumbers(in=t);
    if s and t;
    by number;


proc print data=sample.samptot;
    title 'All Strata';

run;
```

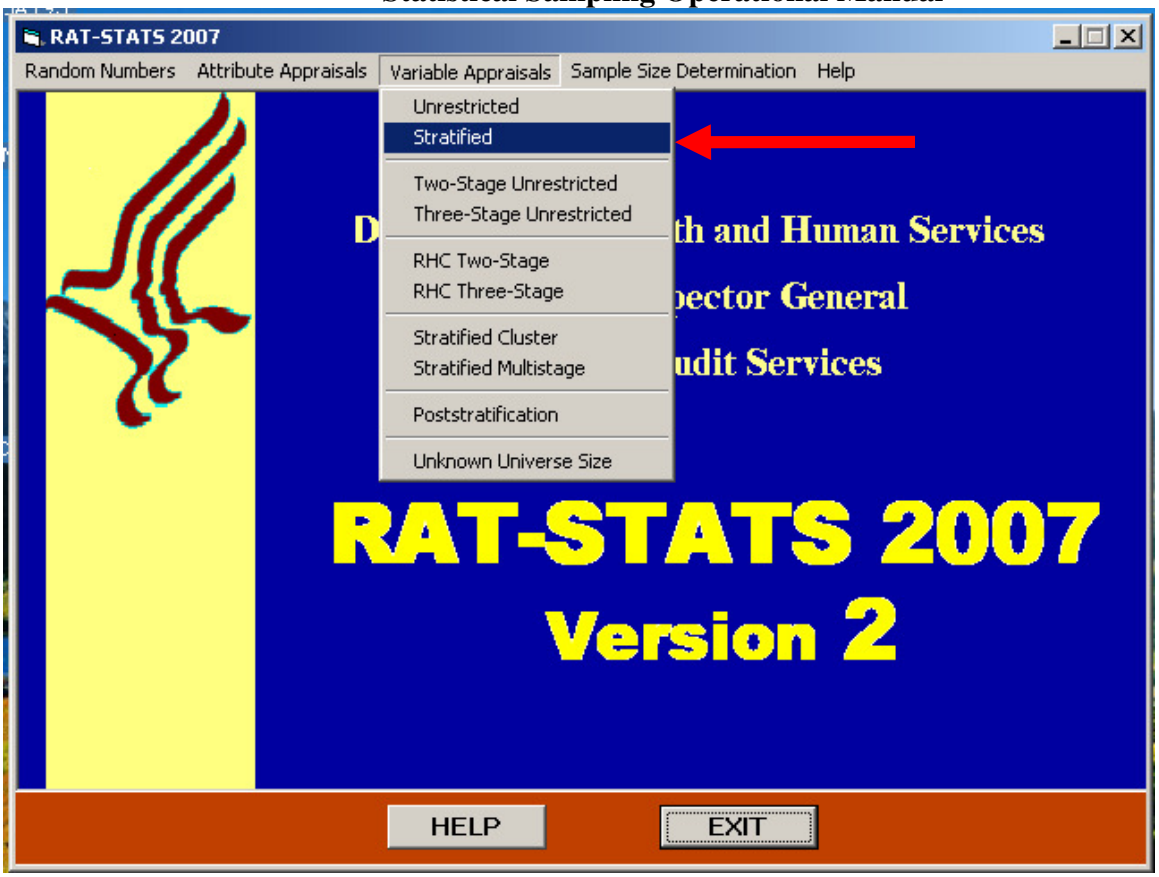
The program numbers each of the records in the 2 strata using the reference table to determine where to start counting for each stratum. It then combines this with the 220 random numbers generated to select those records where there is a match. The output is a sample of 220 records. This is then saved as an EXCEL spreadsheet and submitted to the requestor.

VIII. Sample Submit: Once the sample is obtained it will be delivered back to the requestor who will then begin the audit process. The statistician will keep a copy of the sample and process of obtaining it for a historical record.

IX. Extrapolation Process: Once the sample has been audited, the requestor will enter the amounts that should have been paid for each record in the sample. This information will then be given to the statistician in an EXCEL spreadsheet. It will contain the paid amount and the amount that should have been paid for each sample record. This will then be used in RAT-STATS to determine the amount of money to collect from the entire dataset.

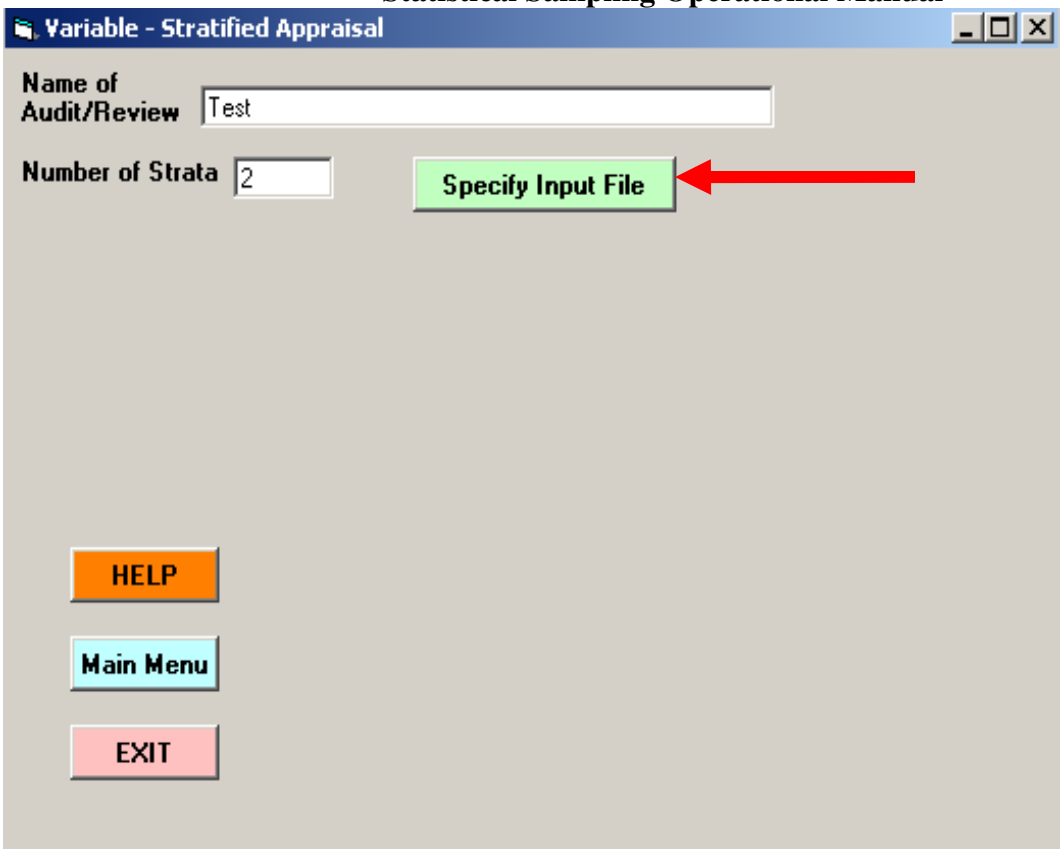
RAT-STATS Extrapolation Process:

Upon entering RAT-STATS again, select the Variable Appraisals, Stratified option.

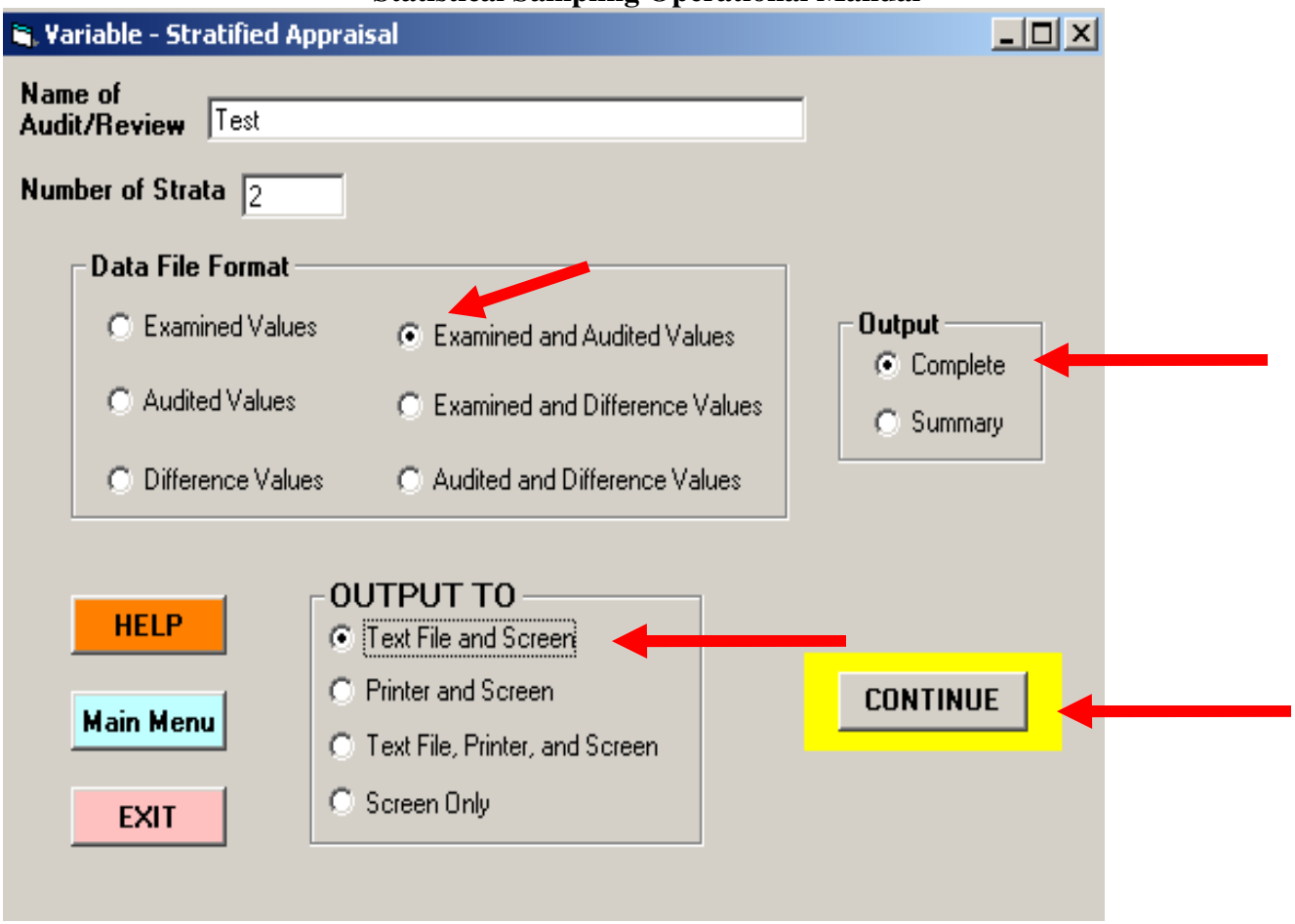


Next enter the name of the Audit, "Test" here, and the number of stratum from the sample, 2 here, and click the Specify Input File button. The input file is an EXCEL workbook that has two spreadsheets. One spreadsheet will contain a column with the total number and sample number of records in each strata. The second worksheet will contain a column with the examined paid amount and the audited paid amount. Both of these are used to determine what the amount should have been for the entire dataset.

Specify here where the input file is saved.

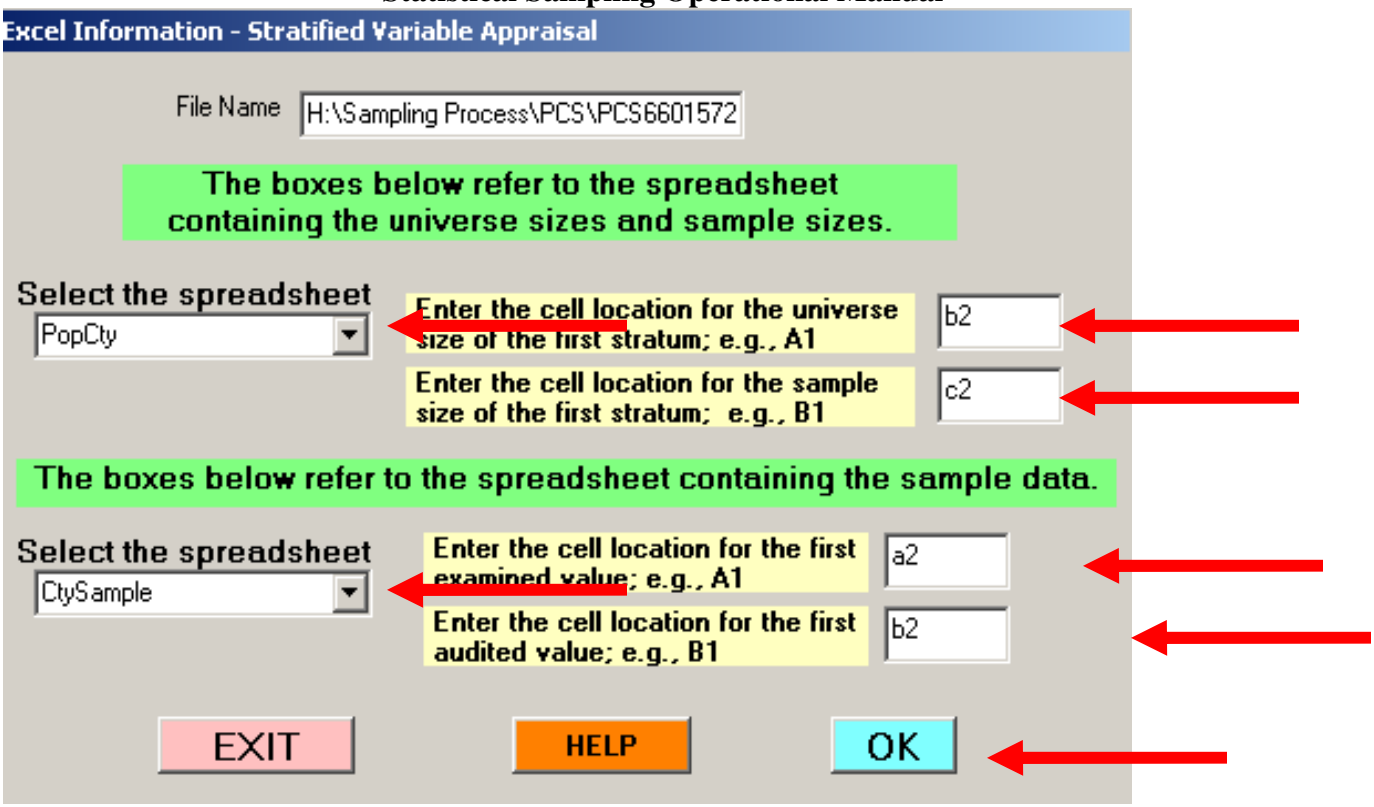


After specifying the input file, the box below will be displayed. Check the Examined and Audited Values circle, the Complete Output circle, the Output To Text File and Screen circle, and the Continue button.

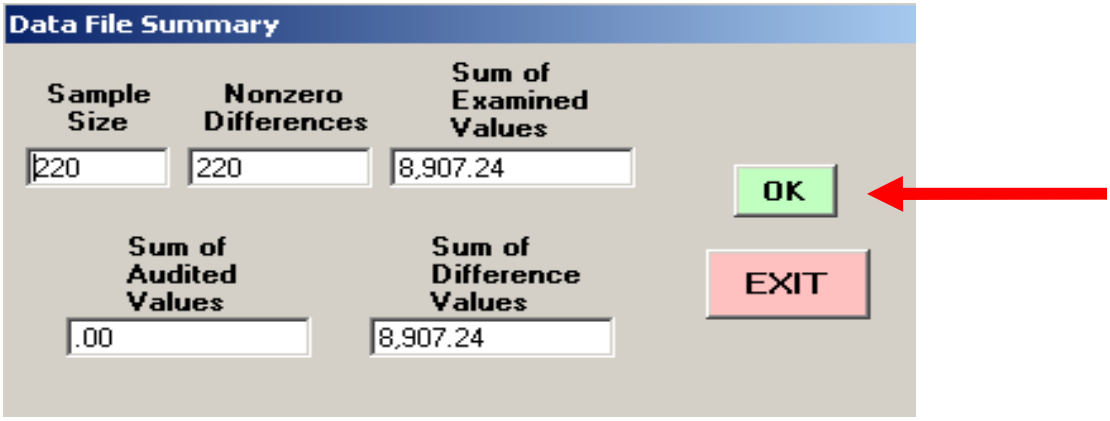


Next, the box below will appear. This is where the locations of the two spreadsheets previously mentioned will be entered. The first spreadsheet indicates the workbook with the total and sample values for each stratum. Indicate which cell in the workbook has the first stratum's value for each.

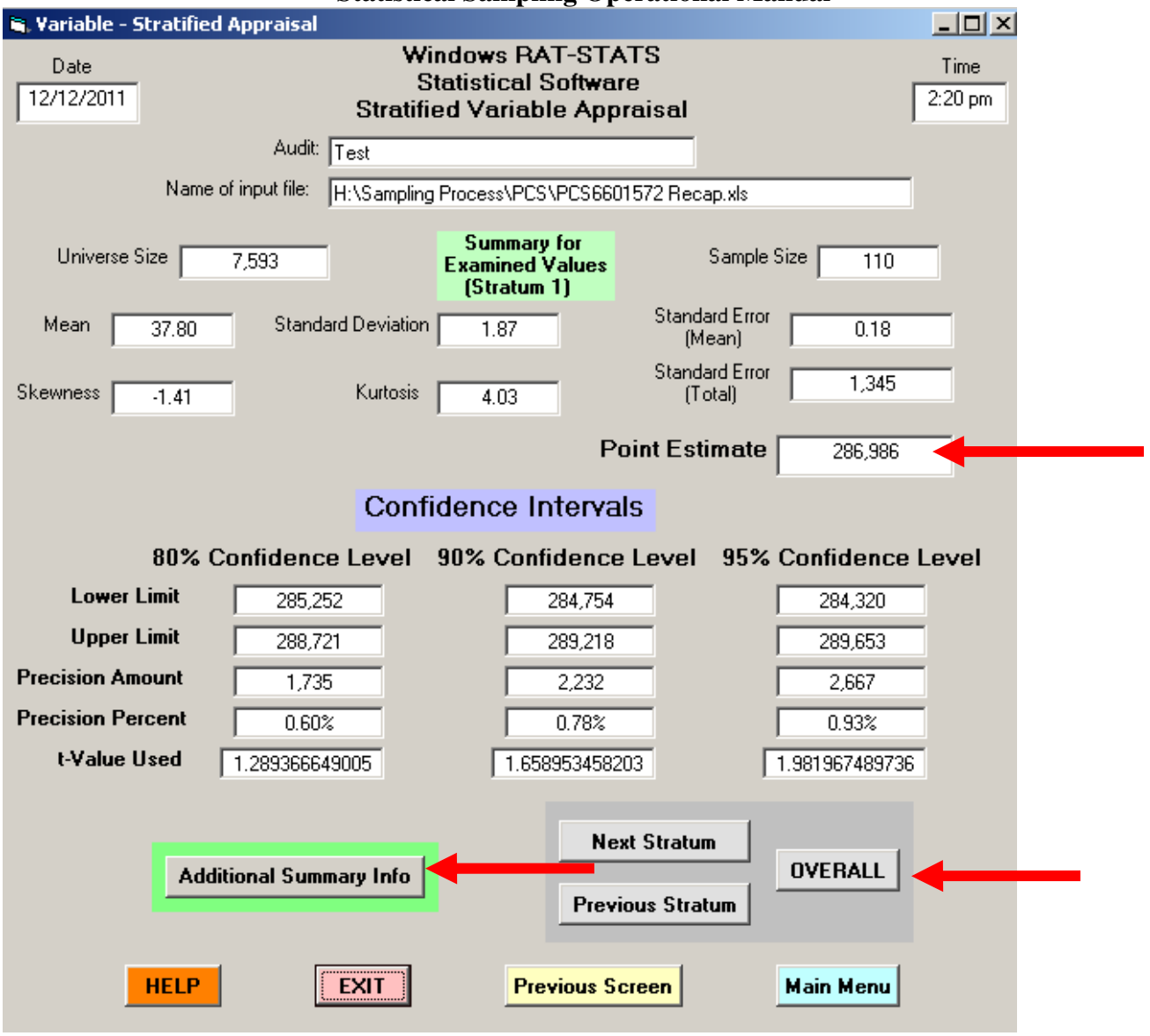
The second spreadsheet indicates the workbook with the examined and audited values from the sample. Indicate which cell in the workbook has the first values for each. Then click the OK button.



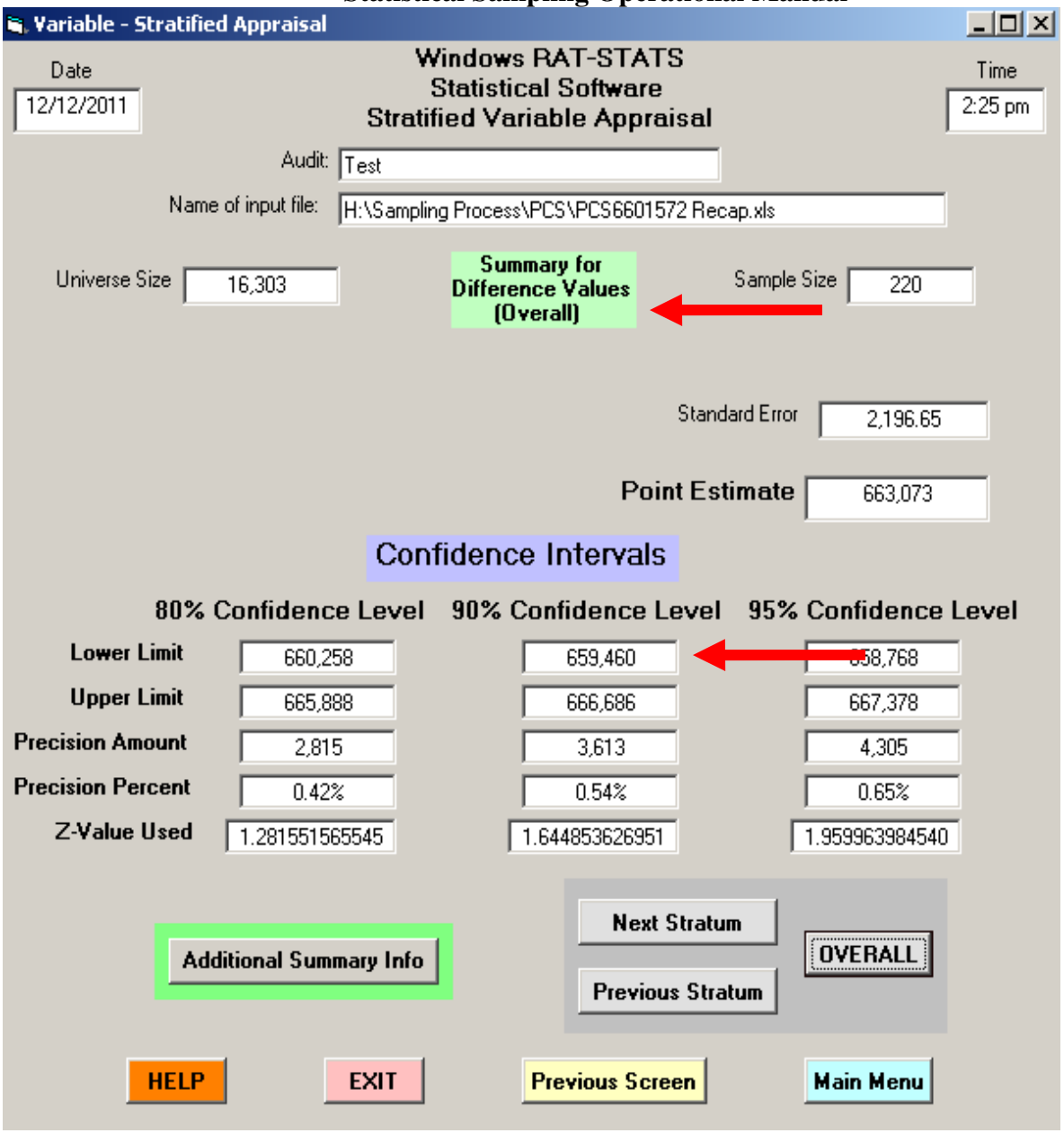
The summary printout below will be displayed:



The screen output will then be displayed. It lists the Examined, Audited, and Difference amounts for each stratum. The Next and Previous Stratum buttons will move between the values for each stratum. By clicking the Additional Summary Info button, the Examined, Audited, or Difference values for each stratum can be displayed.



The amount needed for extrapolation purposes is the overall **90% lower bound Difference value**. Click Additional Summary Info until "Summary for Difference Values" is displayed in the green box. Then click Overall. This is the amount that should be recouped for the entire dataset we are analyzing. It is the amount that should be refunded. In this case it would be $659,460.



The final step is to enter the lower 90% Difference amount into the summary EXCEL spreadsheet that has been referenced throughout the process. Then return this information to the requestor, save a copy for historical record, and the process is complete.

Strata	Mean	Std Dev	N	Sum	Sample	SMean	SSt.Dev.	SSum	Start	End	Start Num
s1 - <40	37.74	2.35	7,593	286,539.66	110	37.80	1.87	4,157.58	1	7,593	1
s2 - >=40	43.46	2.19	8,710	378,510.78	110	43.18	2.10	4,749.66	7,594	16,303	7,594
Sum	40.79	3.64	16,303	665,050.44	220	40.49	3.35	8,907.24			
Recoupment using lower 90% Difference Amount=					$659,460.00						